数字人文与清史研究

地方历史文献的数字化、数据化与文本挖掘：以《中国地方历史文献数据库》为例

赵思渊

［摘要］历史文献数据库可区分为数字化、数据化、文本挖掘三种不同形态，迄今多数中文历史文献数据库实现了数字化功能，部分地实现数据化功能，而能够实现文本挖掘功能的则十分少见。数字化是将文献的物理形态转化为电子形态，数据化是将文献转化为可量化分析的数据，编制元数据是主要方法。文本发掘是在此基础上开发文本分析工具。《中国地方历史文献数据库》以文献学研究为基础，建立特定的元数据结构，提供交叉导航、数据统计等多种功能，这些功能不仅可以帮助研究者找到自己的所需文献，更可能帮助研究者发现新的研究议题。史学研究中，数据库有必要被视作一种新的文献形态，建立针对性的文献学方法论。

［关键词］地方历史文献　数字人文　文本挖掘　元数据
［中图分类号］K249　［文献标识码］A　［文章编号］1002－8587（2016）－04－0026－10

引言

数字化（digitalization）、数据化（datalization）、文本挖掘（text mining）是历史文献数据库的三种不同形态。数字化是将文献从物理形态转化为电子形态，数据化是将电子形态进一步转换为可识别的文本与可分析的数据，文本挖掘则是针对文本、数据做进一步的计量、相关性、GIS 分析。本文将尝试提出并解释 3 种数据库形态分类的依据，并以《中国地方历史文献数据库》为例说明如何实现数据化与文本挖掘。最后，本文将提出一个有待解决的问题，史学研究中，数据库是否已经有必要视作一种新的文献形态，并建立针对性的文献学方法论？

2012 年以来《中国地方历史文献数据库》由上海交通大学出版社、图书馆、历史系合作开发。该数据库由上海交通大学历史系收集资料，并提出数据库建设构想，于 2012－2013 年间由上海交大图书馆进行文献整理与数据加工[①]，2013 年以来由上海交大出版社进行数据库研发。

［收稿日期］2016－05－20
［作者简介］赵思渊（1985—），男，上海交通大学人文学院历史系讲师，上海 200240；titaner@ sjtu. edu. cn
［基金项目］本文系上海市晨光计划“十九世纪徽州乡村的土地市场与社会关系网络”（项目编号：14CGA013）阶段性成果。

① 李芳、陈进、王昕：《上海交通大学新藏地方历史文献的数字化建设规划与实践》，《大学图书馆学报》2015 年第 2 期。

该数据库主要收录上海交通大学 2009 年以来陆续收集的浙江、安徽、福建等地地方历史文献及 2007 年以来曹树基收集、授权复制的《石仓契约》，总计近 35 万件，目前已进入数据库的有 10 万余件。

一、从数据化到文本挖掘：历史文献数据库的演进

数字化并非一个新鲜概念，通常语境中，数字化是指将文献的物理形态转化为电子形态，或者说将模拟数据转换为二进制数据。但电子形态的文献除了易于传播外，并不能增强文献的利用价值。如今天广泛传播于网络的书籍扫描电子档，对读者来说，只是将阅读载体从纸本书变成了电脑屏幕，并未真正改变使用者利用文献的方式。如果将“大数据时代”理解为书籍电子档横行的时代，则远不能视为历史学研究的重大变革。

真正能够改变文献利用方式的是数据化，也即将文献转化为可制表分析的量化形式。[①] 历史文献中包含的产量、价格等信息，可以被转换为量化数据，其他描述性的信息，也应通过某种形式转换为可量化分析的数据，这是历史文献数据化的理想状态。目前在针对历史文献的研究方法中，常用的是词频分析、GIS 以及关系网络分析等。

数据化的意义是将利用文献的方式从“读”转变为“分析”，其核心方法是重组文献内容，置入使用者所建立的新的文本或数据结构中，也即文献的结构化。历史学研究中，这也并非新鲜事物。计算机出现之前，史学研究者已经在制作史料编年、人物关系表，经济史与社会史研究中也早已整理了各种数据序列。如何炳棣研究科举与中国社会流动性的关系，在朱卷基础上建立的庞大数据表并未借助计算机技术。计算机技术带来的革新使得研究者可以更高效率地建立并使用数据。如王业键主持建立的《清代粮价资料库》[②]，该数据库建成于 2008 年，最初依靠代码表查询数据，现在已可利用下拉列表查询时间、地域、粮别，是中国经济史研究的基础性数据。

给历史学带来真正深刻变革的是，计算机技术提供了分析数据化文献的复杂工具。文本挖掘的理念，正是由此兴起。从数据化到文本挖掘的演进，以“数字人文”（digital humanity）概念的兴起为标志。这一理念引导了数据库建设、开发思路的转变，人文学研究者不再是被动选择既有的数据库，而是参与数据库建设过程，由其自身研究需要引导数据库开发，数据库开发过程也就成为其研究的一部分。

“数字人文”概念在 20 世纪 90 年代兴起，逐渐取代 20 世纪 70 年代以来的“人文电算”（humanity computing）概念，成为一个增长迅速的交叉研究领域。项洁、王晓光等已经先后梳理了数字人文概念在西方学界的发展及其在中文人文研究中的适用性。[③] 笔者认为数字人文研究还可细分为文献数据库、线上博物馆、网络文本（cyber born text）分析三个子领域。

总的来看，西方数字人文研究更多的力量集中于对网络文本的分析。[④] 互联网出现以来所积累的各类型线上数据，数量已经十分庞大，自然成为研究者们关注的焦点。而且这类研究直接与互联网经济相关，很多不同学科的学者都有兴趣介入。线上博物馆所关心的则是如何将传统艺术领域的“展示”转变为在线的、可视化的、互动的。

---

① 维克托·迈尔·舍恩伯格：《大数据时代：生活、工作与思维的大变革》，周涛译，浙江人民出版社，第 104 页。

② 《清代粮价资料库》，http：//mhdb. mh. sinica. edu. tw/foodprice，发布日期：2014，访问时间：2016 - 07 - 26。

③ 项洁、陈丽华：《数位人文——学科对话与融合的新领域》，项洁编：《数位人文研究与技艺》，台湾大学出版中心，第 9—23 页；王晓光：《“数字人文”的产生、发展与前沿》，《方法创新与哲学社会科学发展》，武汉大学出版社，2010 年，第 207—221 页。

④ David M. Berry ed. , *Understanding Digital Humanities*, Palgrave Macmillan, 2012, p. 4.

相较而言，历史文献虽然留存数量庞大，但已经电子化的规模仍远远少于网络文本，并且历史文献数据库研究的收益回报也显然低于网络文本研究。因此，针对历史文献的数字人文研究并没有如网络文本那样活跃。虽然如此，如前所述，不论在西方学界或中文学界，都已经有很多数字人文导向的历史文献数字化或数据分析研究。今后的历史文献数字化过程中，数字人文导向将是一个总的趋势。

近年欧洲史研究中已经出现越来越多以文本挖掘为主要目的的数据库或分析工具。如 Tara Andrews 开发的拜占庭文书校勘（critical editing）工具。① 此外，荷兰、比利时等国家在 2013 年集中讨论了“大数据” （big data）对历史学研究的影响，他们所开发的 Biland 数据库以及 WAHSP 数据库可以对 17—18 世纪欧洲的媒体资料进行词频分析、语言比较分析，为人文学者提供帮助。②

中国史研究中，21 世纪初时已有不少学者开始考虑如何使用数据库便利文献检索与研究。③ 这些讨论中，多数学者的关注点是如何使用数据库，而较少涉及如何开发针对性的数据库，研究者参与数据库设计、开发者更少。这一时期代表性的中国史史料数据库是《文渊阁四库全书》电子版与《中国基本古籍库》。《文渊阁四库全书》电子版由上海人民出版社与香港迪志文化公司、香港中文大学共同开发，于 1999 年投入市场。④《中国基本古籍库》于 1998 年作为高校古委会项目立项，由北京大学领衔开发，完成于 2001 年，此后陆续投入市场。⑤

作为第一代中文史料数据库，当时的主要技术难点是文字录入与标准化，实际也就是数据化问题。《四库全书》电子版在研发中曾与清华大学计算机系合作，开发“多特定人准规范手写 OCR 引擎”，用于文字自动录入。如何处理 Unicode 字符集之外的文字，以及如何利用 XML 语言建立文字标引，在当时都是有待解决的技术难题。⑥ 传统文献学中的版本考辨，也是这类数据库所面临的困境，在当时的开发条件下并没有很好地解决。此外，《四库全书》电子版与《中国基本古籍库》最初都使用光盘版发行，这是由当时的技术条件与网络速度决定的。

因此，以上数据库所体现的设计理念是将其视作检索、获得文献文本的储存平台。尽管当时的研究者已经认为“电子版不是纸版翻印”，应当具有丰富的研究功能与工具，但是他们所指的研究功能主要还是检索功能。

2007 年以来，历史文献数字化的范围扩大到古籍以外。有越来越多学者讨论民间历史文献、地方历史文献数据库，除历史学者外，也有图书馆学学者基于各图书馆的馆藏情况，提出特色文献数据库建设构想。但不论怎样讨论历史文献数据库的使用或建设，多数研究者构想的主要是数据库的资料储存、文本检索功能，而较少考虑如何使用数据库中帮助研究者分析文本。

中文民间文书、地方文献数据库中，迄今文本分析、数据处理功能最为完备的是台湾历史数字图书馆（THDL），该数据库由项洁领导的台湾大学数字人文研究中心开发，主要收录台湾地区契约文书及台湾总督府档案。THDL 中提供词频分析、上下手契关联分析、人物相关性分析等

---

① Tara Andrews, “The third way: philology and critical edition in the data age”, working paper, in Lectio Round Table *Digital or critical/digital and critical*?, Leuven, 2011.

② Joris van Eijnatten, Toine Pieters, Jaap Verheul: “Big Data for Global History: The Transformative Promise of Digital Humanities”, *Low Countries Historical Review*, 2013, 128（4）: pp. 55—77.

③ 包伟民:《论当前计算机信息技术对传统历史学的影响》,《杭州大学学报》1998 年第 2 期；王文涛:《古籍数字资料应用与史学研究》,《史学月刊》2009 年第 1 期；陈鹏:《新世纪以来的史料型数据库建设与中国近代史研究》,《国家图书馆学刊》2013 年 6 期。

④ 程之:《香港推出〈文渊阁四库全书电子版〉》,《出版参考》1999 年第 16 期。

⑤《中国基本古籍库光盘工程基本完成》,《图书馆理论与实践》2001 年第 2 期。

⑥ 张轴材:《〈四库全书〉电子版工程与中文信息技术》,《电子出版》1999 年第 3 期；朱岩:《谈古籍数位化》，澳门图书馆编:《“两岸三地古籍与地方文献”会议论文集》，澳门图书馆，2002 年，第 143—150 页。

不同功能，还可以部分地实现契约空间分布的展示。THDL 提出了数据库建设的新理念，那就是数据库的主要功能是为研究者提供研究环境并帮助研究者发现问题，而非仅仅是储存与检索。①

由上可见，迄今流行于网络中的中文历史文献电子资源中，数量最庞大的是扫描、录文、影像等数字化资源，如大量的书籍扫描电子档，以及部分全文检索数据库。此外借助计算机技术实现的文献数据化成果正在逐步积累，如王业键主编的《清代粮价数据库》。数据化基础之上，文本挖掘的发展还比较有限，其代表是台湾历史数位图书馆。

数字人文导向，提供文本挖掘能力将是今后历史文献数据库开发的大趋势。但是，如何数据化？研发怎样的工具能够实现文本挖掘？中文史料数字化的进程中，以上问题还尚在探索之中，成熟的案例并不多。因此，我们在开发《中国地方历史文献数据库》时，将以上问题作为我们的研究焦点。

二、基于文献性质的数据库结构与分析工具研发

我们在开发《中国地方历史文献数据库》的过程中意识到，对文献进行有效的数据化，并开发有效的分析工具，必须以对文献性质的深入研究为基础。数据库开发中，我们主要面临两个问题，第一，如何针对地方历史文献的文献性质，进行有效的数据化。数据化不仅仅是文字录入，更重要的是为文献设计元数据（metadata）。利用元数据标引并标准化文献中的信息，才有可能将文献中的描述内容转变为可分析的数据。

第二，如何从数字人文的理念出发，开发更多有助于研究者的分析工具。今天计算机技术能够提供的分析功能非常多，但不同的软件、分析工具，都对数据类型有特定的要求，因此需要考虑特定的文献类型可以被处理成怎样的数据形态，并据此做针对性的分析工具开发。为了解决这两个问题，首先必须对地方历史文献的性质做一分析。

本文所讨论的地方历史文献，主要指两类材料，一类文献是留存于民间，产生于民间的日常生活，以手写为主，未经过出版暨知识再整理的过程，也可称之为民间历史文献或民间文书。②另一类文献是由地方政府形成的各种档案。这里所说的地方政府主要指作为“亲民之官”的县级或次县级行政机构，对于明清时代来说，也可包含府（州、厅）级行政机构。如民国时期江津县保留了 2 万余卷司法诉讼档案，通过这些档案，可对 20 世纪上半叶的江津地方社会做深入研究。

这类材料与一般意义上的“古籍”具有不同的文献学特征。古籍是经过有意识的书写与知识再组织之后形成的，地方历史文献的文本形成之后，没有经过一个知识再组织的过程，这些文本经历了一个功能性使用的周期后，就被以其使用中的形态保存起来。这意味着，首先，这类文献的每一件都是独一无二的，几乎没有复本。进而，由于没有复本并且未经过知识再组织，这类文本并不形成版本，古籍则具有抽象概念的“书”与作为实体的“版本”之间的分离。③ 这就意味着整理地方历史文献时，版本整理、校勘不是主要难点。

地方历史文献与古籍的另一个差异是，地方历史文献更多情况是碎片化的，单个文本的字数少，古籍整理中所注重的文本内关系，如篇章顺序、自校等，在地方历史文献中虽然也存在，但

---

① 涂丰恩、杜协昌、陈诗沛、何浩洋、项洁：《当资讯科技遇到史料——台湾历史数位图书馆中的未解问题》，项洁编：《数位人文研究的新视野：基础与想象》，台湾大学出版中心，2011 年，第 21—44 页；项洁、翁稷安：《数位人文和历史研究》，项洁编：《数位人文在历史学研究的应用》，台湾大学出版中心，2011 年，第 11—20 页。

② 郑振满：《民间历史文献与民间文化传承研究》，《东南学术》2004 年第 1 期；梁勇、郑振满、郑莉：《新史料与新史学——郑振满教授访谈》，《学术月刊》2012 年第 4 期。

③ 乔秀岩：《古籍整理的理论与实践》，《版本目录学研究》第 1 辑，国家图书馆出版社，2009 年，第 7 页。

不是非常显著。整理地方历史文献时更注重文献之间的关系，以明清史学界整理过程最久的徽州文书为例，以下学者们所提出的徽州文书特性，或可启发我们理解地方历史文献的特性。

表 1　徽州文书特性归纳

周绍泉	臼井佐知子	中岛乐章	严桂夫 王国键	刘伯山
具体性	史料丰富	原始性	系统完成	唯一性
连续性		时间跨度大	年代跨度大	连续性
启发性		数量大	数量大	
真实性		史料丰富		
典型性				

资料来源：周绍泉：《徽州文书与徽学》，《历史研究》，2000 年第 1 期；臼井佐知子：《徽州文书と徽州研究》，載森正夫编：《明清時代史の基本問題》，汲古書院，1997；中岛乐章著、郭万平、高飞译：《明代乡村纠纷与秩序：以徽州文书为中心》，南京：江苏人民出版社，2006；严桂夫、王国键：《徽州文书档案的特点与价值》，《档案学研究》，2001 年第 1 期；刘伯山：《徽州文书的遗存及特点》，《历史档案》，2004 年第 1 期。

周绍泉认为徽州文书具有真实性，因为徽州文书是从实际生活中直接形成的文件。他所说的典型性则是指利用徽州文书可以形成一个个具有代表性的个案研究。中岛乐章所说的原始性，其含义接近与周绍泉所述的真实性，特别强调徽州文书来自实际生活。另外，中岛乐章所说的丰富性是指："徽州学研究的最大优势在于，以徽州文书为中心，大量地保存了长时期族谱等文献史料和建筑等非文献史料。……有可能恢复包括民众文化、日常生活在内的一个地方社会的全貌。"① 严桂夫和王国键所说的系统完整，与刘伯山所述的连续性具有相近含义，均强调徽州文书的来源是可追溯的，文书之间的内部联系是有机的，可以复原的。

以上各位代表性学者所提出的徽州文书特性，可以归纳为以下共同点：第一，所有学者都认为徽州文书存量之大，内容之丰富，是同时代其他文献群难以匹敌的。第二，相对于传世文献，徽州文书的特别之处是其保持了原始记录，同时具有完整的、有机的文献内部联系。

地方历史文献的单件当然也具有研究价值。以契约文书为例，傅衣凌、章有义、杨国桢等前辈学者都曾依据一件件独立的、经过选择的契约解释明清乡村的地权结构。但随着研究的深入，对单件文书的分析、考释，常常不能满足研究的需要，即使在傅衣凌开创契约文书研究的时期，当他对契约文书内容和类型进行了解释和考释之后，也转入了以时间、地域等关系对多件契约做综合分析的研究。可以说，地方历史文献碎片化的形态决定了其每一个单件的研究价值通常要置于一个整体中才能被发现，即其研究应当以一个"文献群"为单位展开。

以上差异决定了，地方历史文献不能使用既有的古籍数据化方法。多数古籍的数据编目，都可参照现代书籍标准。但在地方历史文献的文献结构中，著作人、出版方、出版地点等等都是不主要的，甚至是不存在的信息。因此，必须设计针对性的元数据方案。

数据库开发实践中，我们参照图书馆界通行的都柏林原则（Dublin Core）设计了事主、题名、时间、地域、文献类型等元数据项目。数据库中元数据格式主要实现两种功能。其一是识别每一件文献，并说明文献的性质，如文献编号、资源类型。其二是对文献内容的描述，地方历史文献所涉及的内容千差万别，设计能够适用于全部文献的元数据是非常困难的。因此元数据的设计必须具有高度的弹性，能够涵纳多数文献，如文献名称、涉及人名（事主）、文献归户、日期等，几乎所有文献中都具备。但另一方面，针对存量特别多的文献，也需要针对性设计。从目前粗略的统计看，契约、账簿占到文献收藏的 60% 左右，因此也设计了如标的、金额等此类材料

① 中岛乐章：《明代乡村纠纷与秩序：以徽州文书为中心》，郭万平、高飞译，江苏人民出版社，2010 年，第 43 页。

所特有的元素。①

从资料中提取元数据可以采用人工与半自动标记（semi-automate tag）甚至全自动的方式。上海交通大学目前采取的是人工编目的方式，但是社会学界及数字人文领域已有一些可应用于中文文献的半自动标记工具②，可以预见，这将成为今后的一个趋势。

三、归户：制度史源流、整理方法、元数据

以上元数据格式中，归户是我们首创的元数据项目。这个元数据项能够帮助使用者感受到文献本来的特性，也是进一步开发分析工具的基础。“归户”元数据项体现了我们提出的基于对文献性质的理解构建元数据结构的数据库开发理念。

之所以提出这项设计，是因为我们面临一个困境：地方历史文献与书籍存在文献性质的差异，其研究价值必须以一个“文献群”为单位，那么，如何确定一个文献群的范围？如何在数据化中体现一个文献群的内在联系？

一个具有研究价值的文献群，应当是一组具有内在逻辑关系的文献所组成的整体，特别是那些由生产自同一个来源的文献所形成的整体，如出自同一个家族的全部文书，或同属一个案卷（record）的全部档案。凡是属于同一个文献群的文献，即使是在研究者看来可能并无价值的残件，整理时都应当全部收录。在《石仓契约》的整理与研究过程中，以上方法被归纳为“有机”的研究方法。③

进而我们发现，档案学中的全宗原则、来源原则对如何界定一个文献群有直接的借鉴意义。如果参照全宗原则与来源原则，来自明清赋役制度以及徽州文书的“归户”概念则是最有效界定文献群的方法。

全宗原则和来源原则是19、20世纪之交档案学逐渐发展出的档案管理理论。16至18世纪的欧洲国家，其档案管理本来依据“事由原则”，即按照档案内容对档案进行分类保管。19世纪之后，本来的王室档案馆与行政机关文件登记室逐渐转变为国家档案馆，并且从封闭保密转为开放查阅，档案来源与档案查阅需求也随之多元化，因而，本来封闭的、依照逻辑进行主题分类的档案管理办法不再能满足需要。有的档案可以归入多个分类，或者有的档案不能按照现有分类归档，这都给档案管理造成困难。

1841年，法国内政部第14号通令颁布省档案馆条理，规定：“来源于一个团体、一个机构、一个家庭或者一个人的所有文件都要组成全宗；档案管理人员不得把全宗拆散或将不同的全宗混在一起。”④ 这一条例规定提出了“尊重全宗原则”（the principle of respect pour les fonds），成为“来源原则”、“全宗原则”之滥觞。

继法国之后，1881年德国国家档案馆发布《国家机密档案馆档案整理条例》，其中提出“国家机密档案馆内文件按其组成部分的来源进行整理”以及“每一机关一旦开始移交文件，就要立即指定一部分库房专放该机关的文件，在这部分库房内，官方文件要保持它在有关机关活动过程中获得的顺序和标志。”即“登记室原则”，这一原则之后发展为档案学中通行“来源原则”。⑤

---

① 张洁、李芳、汤萌：《契约文书描述性元数据规范设计与应用》，未刊稿。

② 何浩洋（Hou Leong Ho）：《MARKUS：中文古籍半自动标记平台》，www. academia. edu，发布：2014－12，访问时间：2015－11－27。

③ 蒋勤：《清代石仓文书的“在地”与“有机”分析》，《上海交通大学学报》2014年第3期。

④ 冯惠玲：《论档案整理理论的演变与发展》，载吴宝康、丁永奎：《当代中国档案学论》，中国档案出版社，1988年，第115—167页。

⑤ 黄霄羽：《魂系历史主义——西方档案学支柱理论发展研究》，中国人民大学出版社，2006年，第35页。

地方历史文献中的每一个文献群，正如同档案学中所说的“来源”。近年民间文书整理中所提出的“归户”概念，与来源原则有相似之处。归户是一个来自明清赋役制度的概念，意指赋役过割至地权买入人户，如清初陆陇其总结地方官的为政经验，“受业之家”即地权买入方应当“割税归户”，这里的“归户”是一个动词，为归入买人户之意。明清之际的赋役制度改革中，“归户”是一个总体性的原则。①

> 夫有田则有赋，顽猾抵官者，诚所当治，而善良乐输者，要当与之覆议。其大要，则于移割宜加意焉。产去税存，不可不察，民又以出业报者，便当关会受业之家，割税归户，然后却、与、除、退，庶几无泛追、无滥罚、无推摊抵捱之弊。②

“归户”在明末演变为一个名词，徽州文书中存在“归户亲供册”、“归户清册”等赋役册籍，通常是一个纳税户所有应纳税粮之土地的登记，与陆陇其所称之“归户”涵义相通。根据目前学界对清代赋役制度的理解，这些纳税户通常是一些虚拟户名，其背后可以是个人、家庭、宗族、会社或其他社会团体。③ 这些“户”是纳税单位，同时也即经济活动的单位，进而也是产生契约、账簿等民间文书的基本单位。

整理、研究民间文书的学术史中，刘伯山较早将“归户”作为一项原则，认为徽州文书具有归户性。④ 他在编辑《徽州文书》时将同属一个家族的文书称为归户文书。此后，越来越多学者将“归户性”视作民间文书的一项重要特性，研究者在整理清水江文书、太行山文书时，也开始重视归户整理的方法。

正如档案学对“来源”的理解越趋复杂，随着文献收集越来越丰富，作为文献收集、整理基本单位的“户”、“归户”也应当具有更丰富的内涵。事实上，早在1962年严中平已经提出一项针对收集工作的建议，希望能够“完整地”收集徽州文书。⑤ 我们认为严中平所说的“完整”已经包含了“归户”的整理原则。正如档案保管从事由分类转向来源分类，保管、整理地方历史文献也应当以文献群或“归户”作为基本单位，从而取代按照内容、年代等进行整理的原则。⑥ 因为这些文献本来是以文“户”为单位产生的，以“户”或文献群为单位进行保管、分类，最能够保持文献内部的有机联系。同时，“户”的所指也应更加丰富，举凡家户、家族、宗族、会社、寺庙等都可成为一“户”。

因此在《中国地方历史文献数据库》中，“归户”成为一个元数据项目，设计为“县+姓氏”的形式，根据收集文献时获得的信息，标注每件文献所属的县份及姓氏，由此反映文献与当地人群之间可能存在的关系。在徽州及浙南等文献脉络更清晰的地方，文献的归户信息还可细化到县以下层级，也即其所属的“都”、“图”、村落等。但县以下行政区划层级几乎每一县均不相同，因此在按照“归户”信息检索的界面中，省去了县以下层级，而在元数据中，则以文字形式保留了这些信息。

为了弥补以上不足，元数据中又增加了“批次”信息，这是收录于《中国地方历史文献数据库》中每一个文献群的编号，这个编号是根据每个文献群入藏的时间制作的，文献群中的每

---

① 刘志伟：《在国家与社会之间：明清广东地区里甲赋役制度与乡村社会（修订版）》，中国人民大学出版社，2010年，第201页。

② 陆陇其：《莅政摘要》卷上第12页，《官箴书集成》第2册，黄山书社，1997年，第628页。

③ 刘志伟：《在国家与社会之间——明清广东地区里甲赋役制度与乡村社会（修订版）》，第197—204页。

④ 刘伯山：《徽州文书的遗存及特点》，《历史档案》2004年第1期。

⑤ 严中平致中央档案馆函（1962年2月6日），安徽省档案馆藏，转引自严桂夫、王国键：《徽州文书档案》，安徽人民出版社，2003年，第11页。

⑥ 民间历史文献整理方法演进的学术史，参见杨培娜、申斌：《走向民间历史文献学——20世纪民间文献搜集整理方法的演进历程》，《中山大学学报》2014年第5期；张侃：《20世纪以来民间文献研究的学理述略》，“第七届民间历史文献论坛”，厦门大学，2015年。

一件文献则在批次号的基础上流水编号。批次号是对文献群物理保存形态的反映。

四、文本挖掘：拓展数据库应用的可能性

研发《中国地方历史文献数据库》时，由于资金与技术的限制，并未设计嵌入数据库的文本挖掘工具，但从数字人文的理念出发，设计了两种检索方式以及两组检得文献分析工具。研究者利用这些工具，就有可能进行进一步的文本挖掘与研究。

数据库提供的基本检索方法是输入任意词在整个数据库中检索，或者以下拉列表方式在题名、事主、归户、事由、分类中用任意词检索，也就是通常文献数据库都具备的普通检索与高级检索。另一种检索方式是多维分类导航，也即利用时间、地域、归户、分类法等方法交叉浏览、检索，寻找文献。对于检索所得文献，数据库提供两种分析工具，一种可以统计检得文献的地域分布、年代排序、类型分布，以及事主统计。另一种工具则可显示检得文献的关联文献，如属于同一批次、同一地域、同一归户或同一事主的文献及其数量。

建立以上检索与分析工具的意义是为研究者提供更好的研究环境。通过检索找到资料库中的资料，是研究者使用数据库的最基本需求，但是，历史学研究不仅需要找到资料，更重要的是发现资料间的关系。历史学研究中，文本记录中的时间、空间、人物无疑是最重要的三组关系，我们试图在数据库中提供相应的功能帮助研究者揭示资料群在这三个方面的关联性。依据时间检索，检得资料时间分布统计正是为发现不同时间形成的资料间的关系而设计的。地域、归户等则是以不同形式分析、呈现资料的空间分布。文本中所有的人物信息则都被登记为事主。

以上功能设计还有可能进一步帮助研究者发现新的资料或新的研究议题。以我们最近的一项研究为例，本来的研究计划中，只是准备分析清代徽州契约中代笔人与买卖双方的亲属关系，在利用事主相关功能检索一批文书中代笔人的身份时，发现在一批契约中的代笔人江振玉同时还是当地编制归户册的册书，由此开始研究清代乡村中同时担任半职业代笔与税收职役的人群。①

并且，编制元数据时提取了文书中的全部人物，也就有可能分析文书所反映的社会网络。仍以上述研究为例。根据歙县枧桥江氏家族所保存的100余份契约，可以建立该家族清代、民国时期的土地交易记录数据库，图1是根据数据库绘制的当地土地交易社会网络。图1中可以观察到300余年间该家族有实力大宗购入土地的主要成员，以及该家族购买土地时主要使用的户名。此外，利用分家书、家谱等其他资料，还可进一步分析图1中与该家族频繁交易的人物身份。研究者由此可以了解当地社会关系网络对土地交易的影响。

五、结论：建立史料数据库的文献学方法

数字人文是最近20年来新兴的交叉学科概念，对历史学研究而言，引入这一概念的意义是促进更多分析工具应用于文献整理与解读。与古籍不同，地方历史文献未经过知识重组，也相对碎片化，更注重文献间的关联性。《中国地方历史文献数据库》针对文献特性设计元数据结构，从而实现对文献的多维度检索。尤其是我们根据文献特性所提出的“归户”数据项，将帮助研究者发掘文献的内在关联。我们不仅期望这些功能设计可以便利研究者寻找史料，更期望以此帮助研究者发现新的研究议题。

元数据是可以被计算机识别的文献描述，制作元数据，也就是将文献资料转化为可被检索、分析的数据的过程，这是将物理形态的史料转化为可分析的数字形态的关键。如何设计元数据结构，很大程度上决定了文献资料可以被如何检索、分析。另一方面，数据库的使用者有必要了解

---

① 赵思渊：《19世纪徽州乡村的土地市场、信用机制与关系网络》，《近代史研究》2015年第3期。

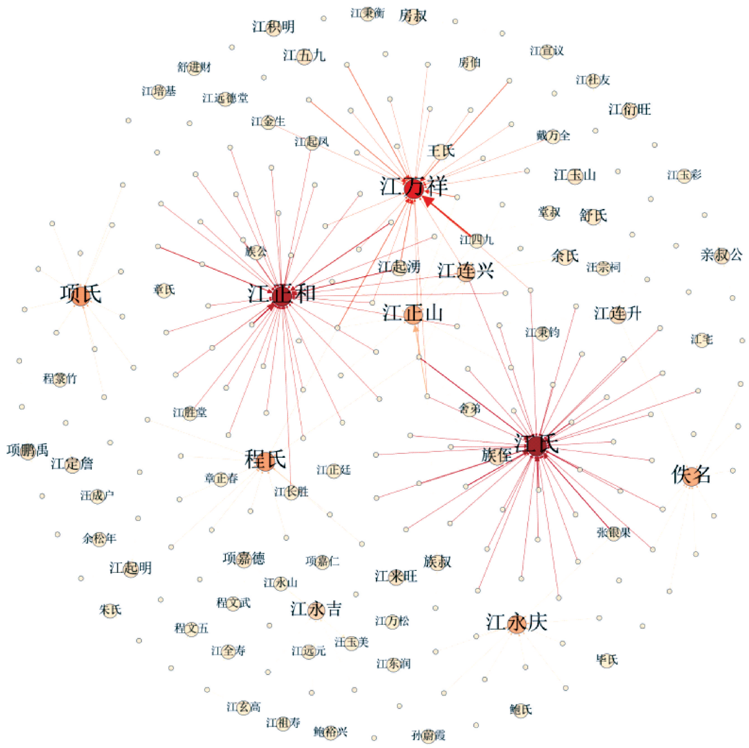

图 1　歙县枧桥江氏地权交易网络（获得地权）

资料来源：《中国地方历史文献数据库》，http：//www. datahistory. cn，批次号：0111120601。

说明：图中以箭头表示土地权利转让的方向，如图中箭头从江四九指向江万祥，表示土地权利从江四九转让至江万祥。图中每个点的颜色表示其在交易中购入地权的次数，频率越高，颜色越深。这里所说的获得地权，包括买入、典入、抵押等形式。

元数据结构，从而判断哪些因素可能影响了自己的分析结果。

因此，历史文献数据库不仅是传统史料的载体或“仓库”，其本身也将日渐形成一种独立的文献形态。历史文献学对传统史料已经形成了一套绵密、精细的处理方法，数据库作为一种新的文献形态也应当建立针对性的文献学方法论。对元数据结构的考辨可能应当是这种方法论的核心。史学理论对史实与史料关系的思考，也同样适用于历史文献与数据库。

References

Huang, Xiao-Yu. *Holding to history creed: General study on the development of the Western Pillar thoery of archives.* Beijing: China Renmin University Press, 2006.

Wang, Xiao-guang. “Digital Humanities: the development and frontier.” In Innovation and Development of Social Science, Wuhan: Wuhan University Press, 2010.

Nakajima, Yoshiaki. *Disputes and order in Rural Society during the Ming Period: An Analysis Based on Huizhou Documents.* Translated by Guo Wanping and Gao Fei. Nanjing: Jiangsu People’ s Publishing, Ltd., 2010.

Liu, Zhi-Wei. Between State and Society: a study of lijia tax and service in Ming and QingGuangdong. Beijing: China Renmin University Press, 2010.

Hsiang, Jie ed. *New eyes for discovery: foundations and imaginations of digital humanities.* Taipei: National Taiwan University Press, 2011.

Hsiang, Jie ed. *Digital Humanities and Historical Research.* Taipei: National Taiwan University Press, 2011.

Mayer-Schönberger, Viktor. *Big Data: A Revolution that Will Transform How We Live, Work, and Think.* Translated by Zhou Tao. Hangzhou: Zhejiang People’ s Press, 2012.

Hsiang, Jie ed. Digital Humanities and Craft: Thecnological Change. Taipei: National Taiwan University Press,

2014.

Luan, Chengxian. "A Study on Qinzhang Guihu Qingongce (Registration on corvee and land tax) in 1581." *Journal of National Museum of China* 2 (1996): 79 – 93.

Bao, Weimin. "On the Influence of Computer Science to the traditional Historical Study." *Journal of Hangzhou University* 28. 2 (1998): 1 – 8.

Zhang, Zhou-cai. "Chinese Infromation Technology in the Project of the Digitalization of Sikuquanshu." *Electronic Publishing* 3 (1999): 3 – 6.

Zheng, Zhen-man. "Archives from Private Holdings and the inheritage of the Folk Culture." *Southeast Academic Research* S1 (2004): 293 – 296.

Wang, Wentao. "Digitizing Ancient Books and Historical Research." *Journal of Historical Science* 1 (2009): 119 – 125.

Qiao, Fu-jin. "Arrangement of Rural Social Literature in the Perspective of Historical Documents Studies." *Journal of Eastern Liaoning University (Social Sciences)* 13. 3 (2011): 103 – 112.

Chen, Peng. "The Historical Database Construction and Research in Modern Chinese History since the New Century." *Journal of the National Library of China* 22. 6 (2013): 33 – 38.

Jiang, Qin. "A 'Contextual-Relational' and 'Organic' Analysis on Local Documents inShicang Village in Zhejiang Province in the Qing Dynasty." *Journal of Shanghai Jiaotong University (Philosophy and Social Sciences)* 22. 3 (2014): 88 – 98.

Zhao, Siyuan and Tang Meng. "Classification of Historical Archives in Private Holding: On the Collections ofShanghai Jiaotong University." *Journal of SYTU (Philosophy and Social Sciences)* 22. 3 (2014): 76 – 87.

Yang, Peina and Shen Bin. "Towards the Science of Folk Historical Documents: the Process of Collection and Editing in 20th Century." *Journal of Sun Yat-Sen University (Social Science Edition)* 54. 5 (2014): 71 – 80.

Li, Fang, Chen Jin and Wang Xin. "Planning and Practice for the Historical Documents Digitalization in Recent Library Holding ofShanghai Jiaotong University." *Journal of Academic Library* 2 (2015): 77 – 83.

Zhang Kan, Analysis on Populace Archives from 20th Century, 7th symposium of Populace Historical Archives, Xiamen University, 2015.

Zhao, Siyuan. "Land Markets, Credit system and Relationship Networks in the 19th Century Rural Huizhou." *Modern Chinese History Studies* 4 (2015): 82 – 97.

Andrews, Tara. *The third way: philology and critical edition in the data age*, working paper, at LECTIO Round Table "Digital or critical/digital and critical?", Leuven, 2011.

Berry, David M. ed. *Understanding Digital Humanities*, Palgrave Macmillan, 2012

Eijnatten, Joris van., and Toine Pieters, Jaap Verheul. "*Big Data for Global History: The Transformative Promise of Digital Humanities.*" *Low Countries Historical Review* 128. 4 (2013): 55 – 77.

The Digitization of Local Historical Archives, Creation of Metadata, and Datamining: The Example of The Chinese Local History Archive

ZHAO Siyuan (History Department, Shanghai Jiao Tong University; titaner@ sjtu. edu. cn)

This article exams three distinct concepts that distinguish historical databases: digitization, creation of metadata, and text mining. Digitization has transformed the archives into cyber texts that encompass data and metadata, which can then be used for text mining, databases, and other analytic tools. The database of Local Chinese Historical Archives was designed with these principles of metadata creation and text mining that constructed a modified metadata based on the Dublin Core. Cross-searching and statistics is also available in the database. The database not only facilitates the researchers' effort to find data for their research but also helps them discover new topics. It is also suggested that a new corresponding methodology on archives criticism should be applied to the database.